Online Grocery Shopping Data

Sava Dashev

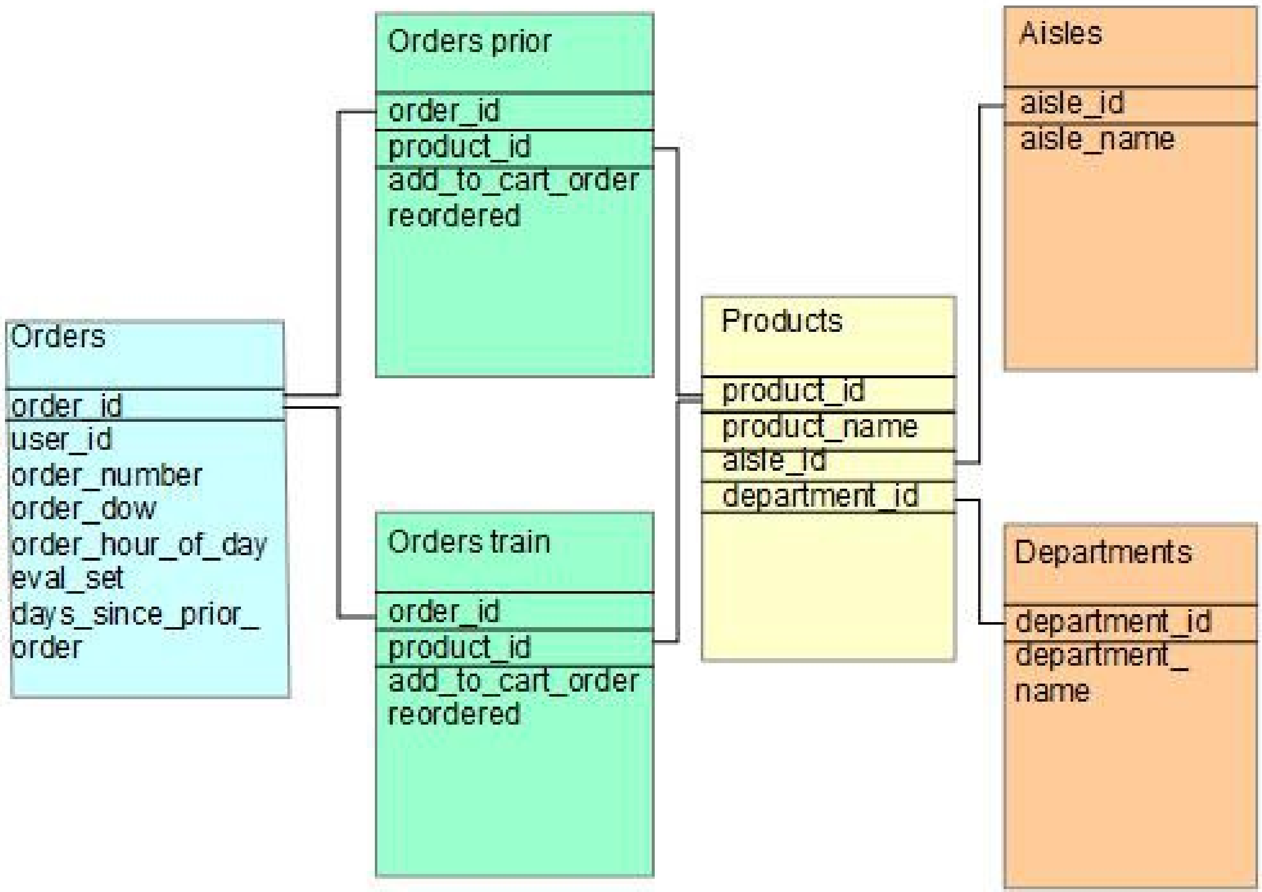
Problem Statement

This survey will try to use unsupervised learning to separate customers into groups

Data set

The Kaggle competition data was collected in 2017.
It is organized as database with 5 separate files.

Data structure



Data Wrangling

The data set is relatively clean.
We check:

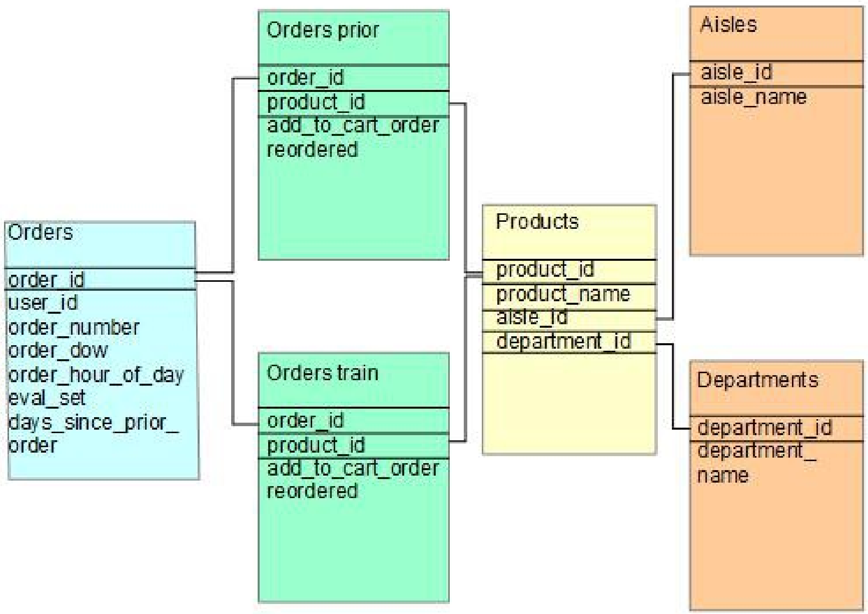
- Completeness;
- Missing values;
- Duplicate orders.

Data Wrangling

We combined data into one large file.

```
1 data_all.count()

order_id                  33819106
user_id                   33819106
eval_set                  33819106
order_number              33819106
order_dow                 33819106
order_hour_of_day         33819106
days_since_prior_order    31741038
add_to_cart_order         33819106
product_id                33819106
reordered                 33819106
product_name              33819106
aisle_id                  33819106
department_id             33819106
aisle                     33819106
department                33819106
dtype: int64
```



Exploratory analysis

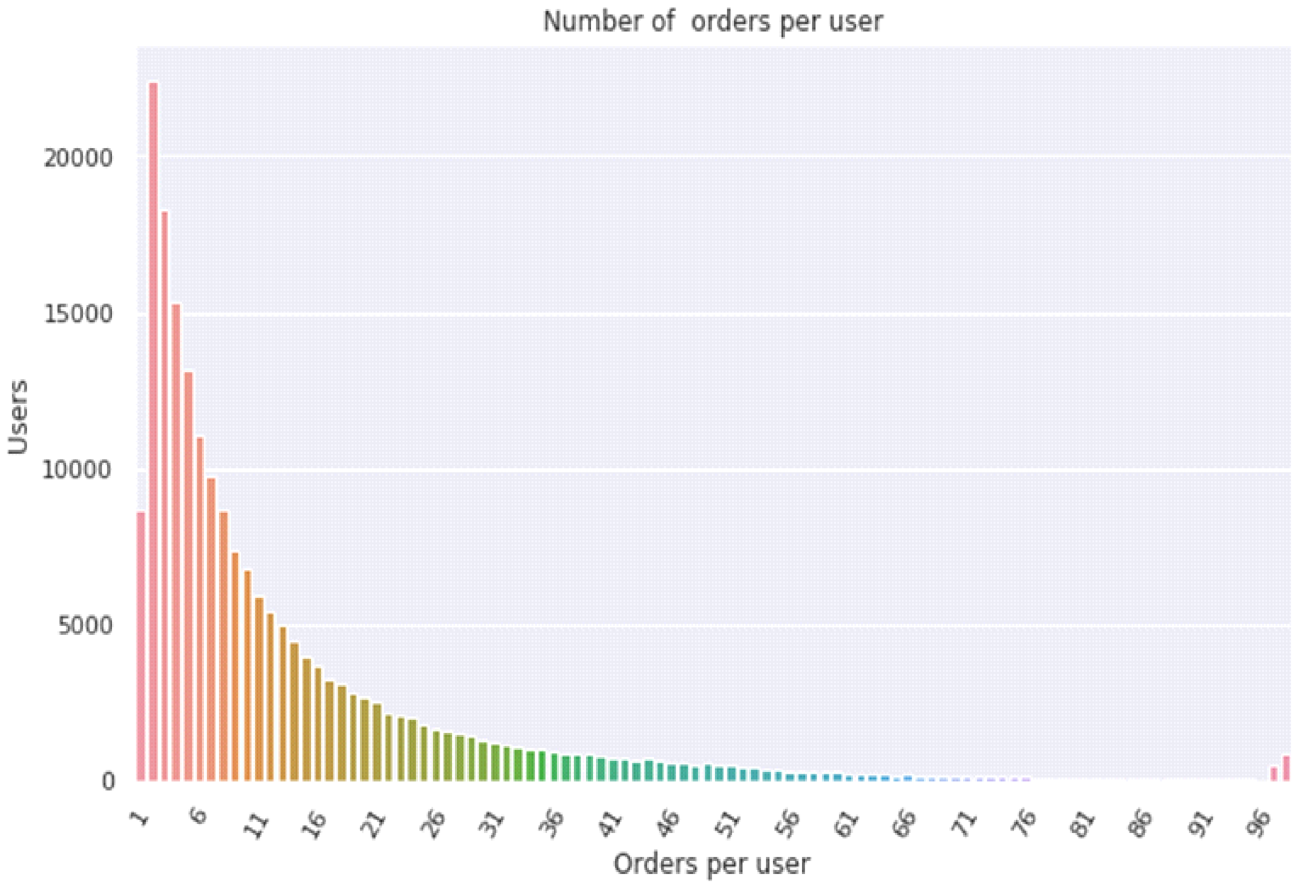
- When are users most and least active?

Exploratory analysis

- Weekly pattern

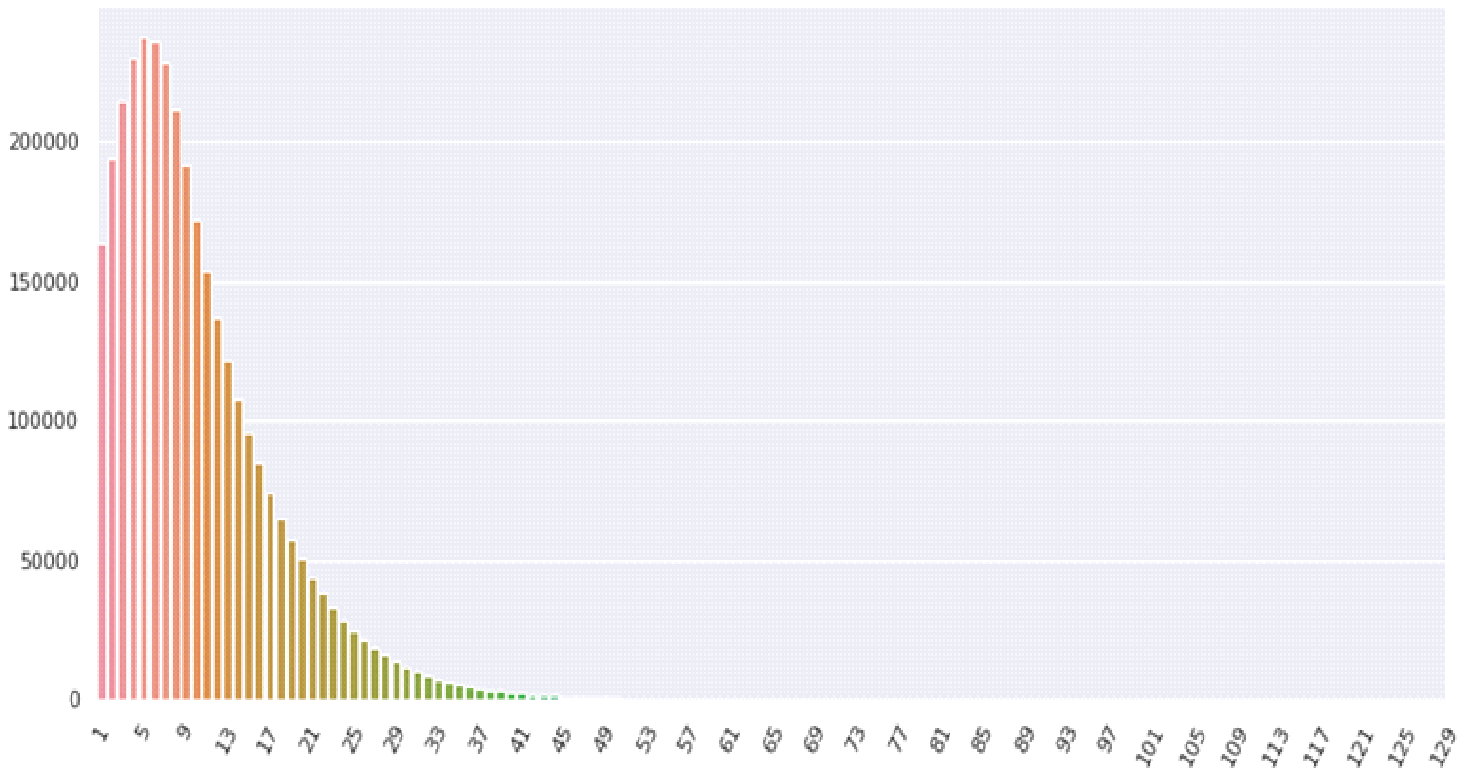
Exploratory analysis

Number of orders per user



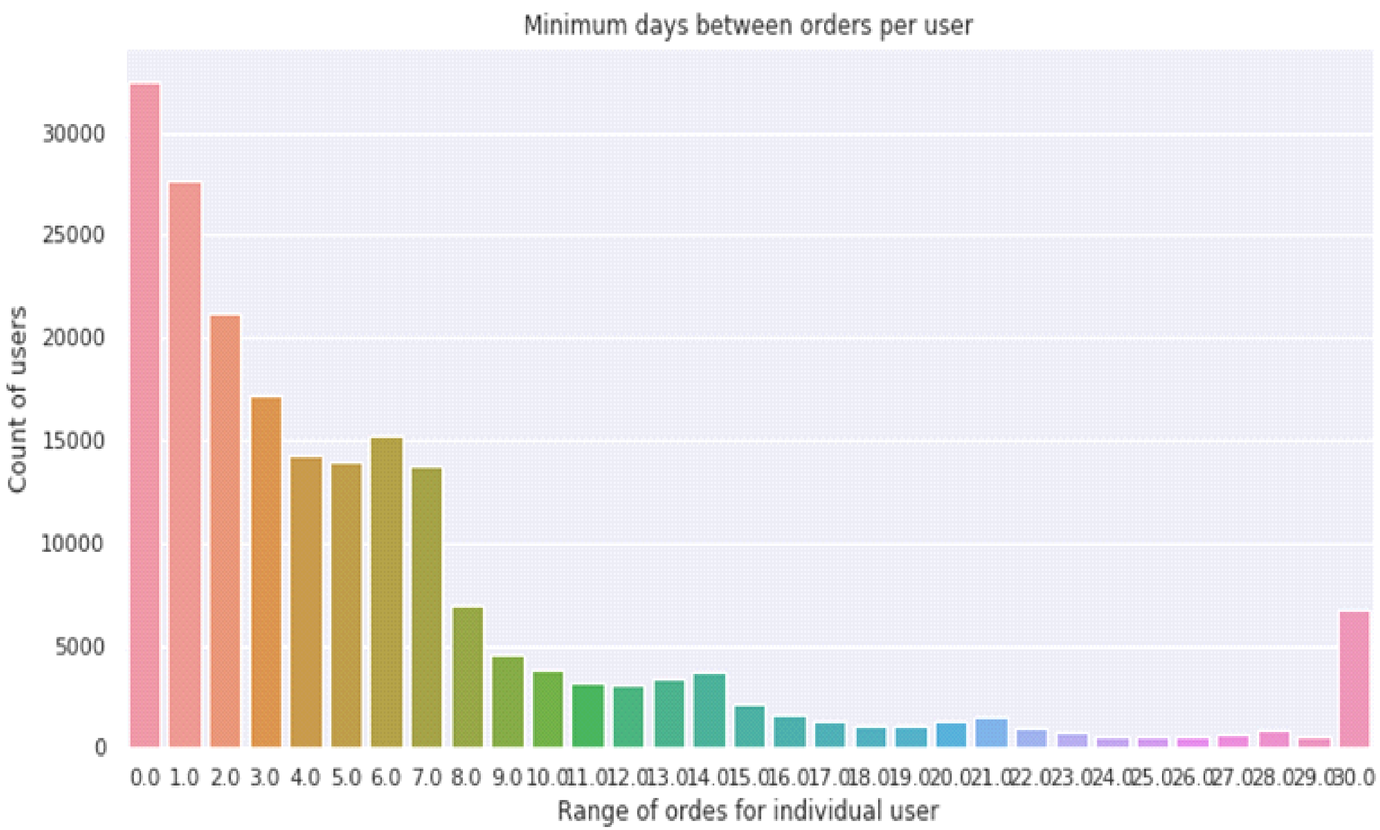
Exploratory analysis

Number of items in order



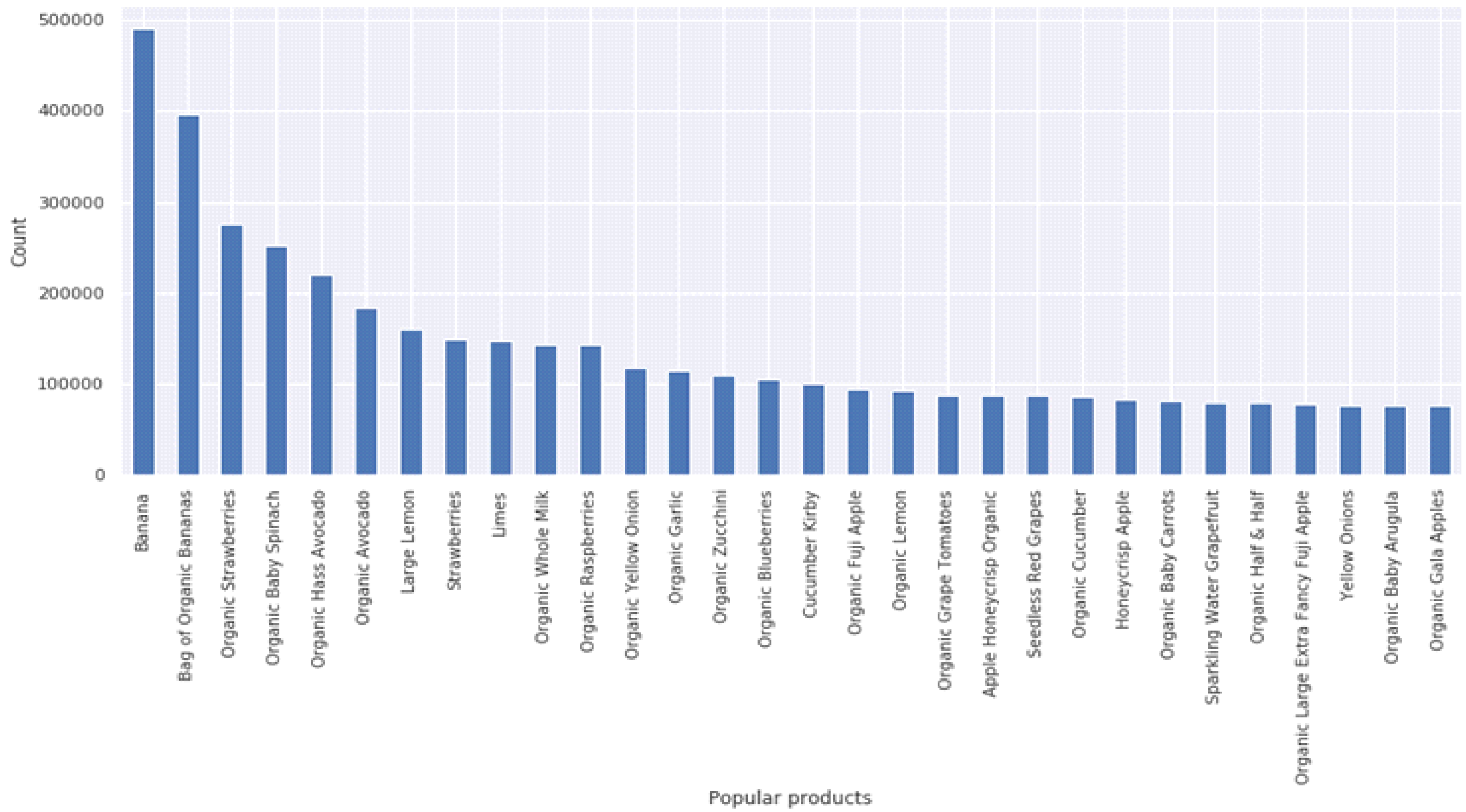
Exploratory analysis

Days between orders



Exploratory analysis

Most ordered items



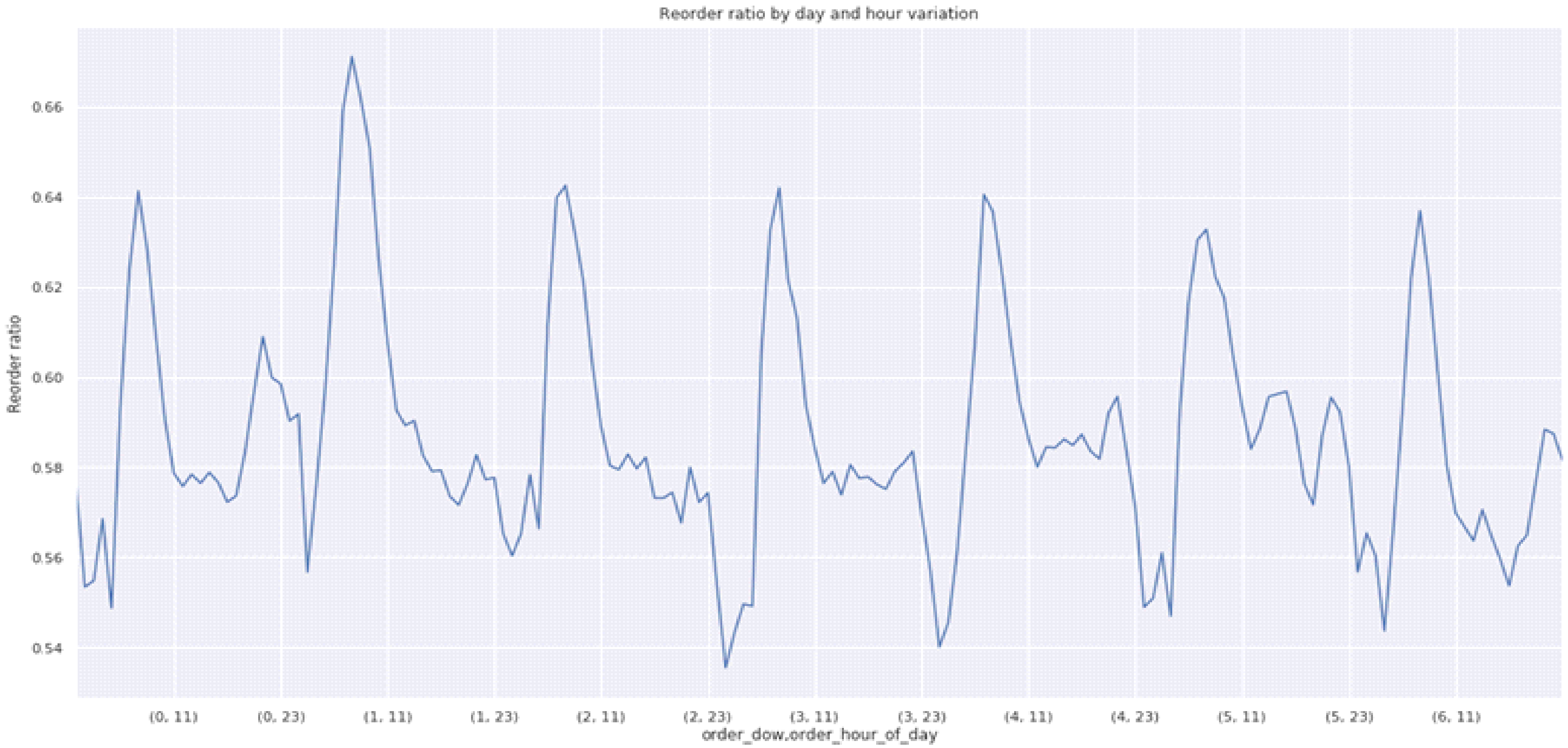
Statistical inferences

Reorder ratio

- Varies by time of day and day;
- Varies by product, aisle and department.

Statistical inferences

Reorder ratio



Statistical inferences

Reorder ratio by product

Reorder ratio by day and time is statistically significant for most of the different days and times.

Statistical inferences

Reorder ratio by product

Reorder ratio by product is statistically significant when the reorder proportion is about 0.15.

Statistical inferences

Reorder ratio by department

Reorder ratio by product is statistically significant, even for department with small number of items and close reorder proprtions.

Machine learning

We used unsupervised learning to divide customers in clusters.
The variable we used to create clusters is aisle.
Going to level of individual products and the number of customers in the base make this approach unfeasible.

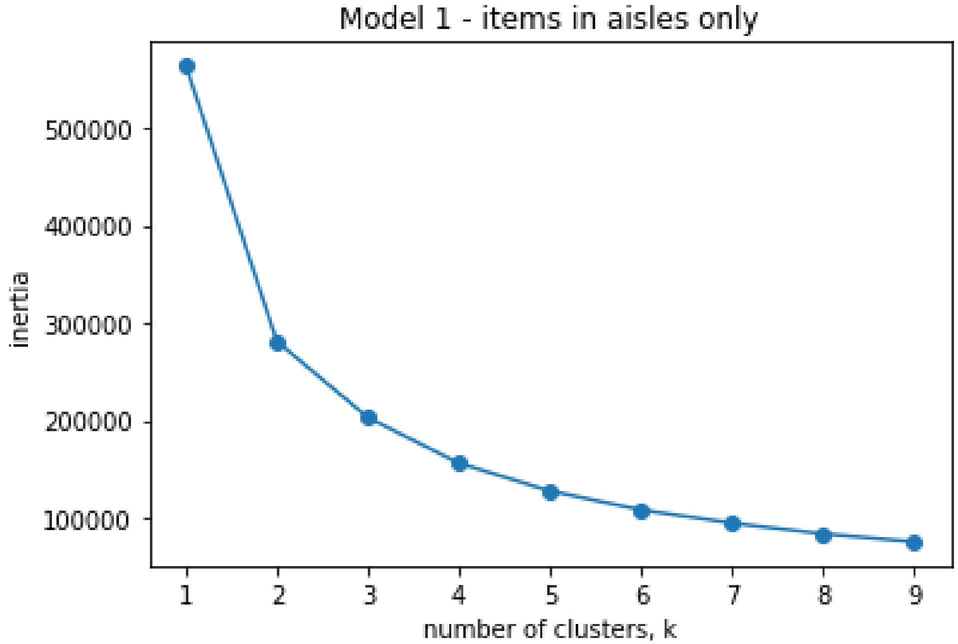
Machine learning

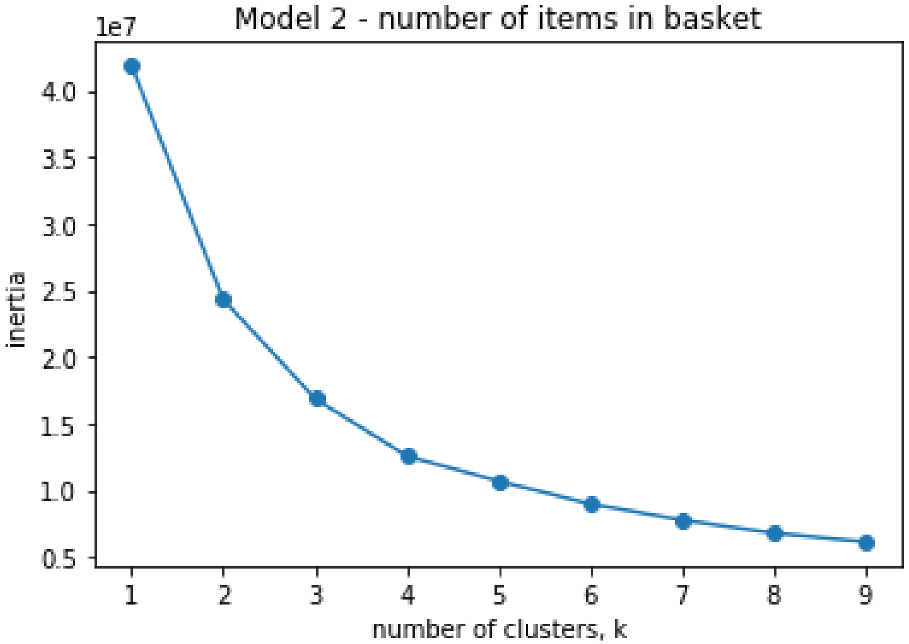
We created three models:
Model 1 – using aggregate data from aisles and customers;
Model 2 – we add the time data and the maximum number of items in basket per user;
Model 2a – we use only the max number of items in basked in addition to the variables in Model 1.

Machine learning

We used inertia as a mean to determine the number of clusters in model.
Model 1 has 2 clusters.
Model 2 has 4 clusters.
Model 2a has 3 or 4 clusters.

Machine learning





Machine learning

Model 2 has the best potential to divide customers into clusters.
Most ordered items are from the same aisles for all customers.
The clusters differ in the second tier of items ordered.

Machine learning

Model 2 has the best potential to divide customers into clusters.
Most ordered items are from the same aisles for all customers.
The clusters differ in the second tier of items ordered.

Conclusion

Customers can be divided into four main groups. Groups differ by number of items in basket, how often they shop, and the ranking by how many items in particular the customers order from a single aisle.

Conclusion

Time is of the essence for this model. We can improve the data by transforming day-time into phase, using 168 hours as a period.

Conclusion

Similar model running on individual products fail. Reducing the data to 5%, the model used all available RAM, and fail.

Conclusion

Some questions posed need further refining the model. We can use only products from aisles in the middle tier.

“The Instacart Online Grocery Shopping Dataset 2017”, Accessed from https://www.instacart.com/datasets/grocery-shopping-2017 on 6-12-2019